From Learners' Corpora to Expert Knowledge Description: Analyzing Prepositions in the NICT JLE (Japanese Learner English) Corpus

Midori TANIMURA[1], Kazuhiro TAKEUCHI[2], Hitoshi ISAHARA[3]
[1] *NICT, 3-5 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0289, Japan*
mtanimura@nict.go.jp
[2] *NICT, 3-5 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0289, Japan*
kazuh@nict.go.jp
[3] *NICT, 3-5 Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0289, Japan*
isahara@nict.go.jp

Abstract

The present study has two main purposes. One is to show what the NICT JLE corpus analysis tool and commercial tools can do for analysing error annotated learner corpora. The other is to show how possible L1 transfer effects can be analyzed using learner corpora. We take preposition errors as examples, which often occur as collocation. For future direction, we suggest that standardization of a shared collocation dictionary rather than a word list is needed, especially for Japanese English learners for pedagogic purposes. We also propose that an objective method to characterize L1 transfer needs to be developed. We assume that back-translation of each utterance would be one of the effective ways to extract L1 translation, in other words, Bi-lingual aligned corpus and machine translation software would be a foundation to develop such a method.

Keywords: *error annotated learner corpora, the NICT corpus analysis tool, preposition errors, collocation*

1 Introduction

Error annotated learner corpora are recent attempts to isolate or codify error-related knowledge. In general, an annotation to a certain error example consists of at least 3 common attributes: incorrect instance, respective corrected instance, and Why-scheme. The third attribute generally describes errors in the following two major ways. Firstly, error-tag based on a predefined error-hierarchy. Secondly, narrative - based on natural language descriptions of the reasons for learners' errors (e.g. "tense sequence error" or "past tense form of *pay* is not *payed*," etc.)

Each attribute provides us with useful information; for example, errors which learners often tend to make and insights into causes of errors. However, for a particular error, the first two attributes, incorrect instance and respective corrected instance, identified by different annotators may not be unique (e.g., N-V agreement error: either N or V can be corrected to achieve conformity). For the third attribute, Why-scheme by error tags, it is easy to compile statistics but difficult to describe errors in detail, since there are various, unpredictable errors occurring in learners' English. The opposite is the case, in looking at Why-scheme by narrative.

In this study, we use Japanese Learner's Spoken English, NICT JLE Corpus (formerly SST corpus, Tono et al., 2001) which contains 1200 files with 9 different proficiency levels (level 1 the most elementary and level 9 the most advanced). Although 167 files have already been error-tagged on the basis of 47 kinds of grammatical terms, there is still room to improve the error tagging system. In order to explore the possibilities of the error-tagged corpus, further investigation into details is required. To this end, we analyze learners' errors focusing in particular on prepositions, which often occur as

collocation even in a low level speaker's utterance (see Tono, 2004), but have not been focused on until now.

The present study aims to
(1) show what the NICT corpus analysis tool can do
(2) show how some other various tools can be integrated to analyse in detail
(3) examine preposition errors
(4) show future direction for collocation standardization and for possible evidence for L1 transfer

This study analyzes the points above to facilitate integration and reuse learner corpora.

2 Procedure

To analyze the preposition errors, this section explains what the NICT JLE corpus analysis tool can do in a practical sense, and how various other tools can be integrated to analyse in detail.

2.1 Preposition error occurrence

The analysis tool is an integrated device for looking at how words behave in text. It allows us to not only search for how grammatical and lexical errors and/or discourse annotations are distributed, but also allows us to go back to the original data and personal information. Lexical frequency, collocations and various types of statistical information can be generated. Moreover, search results are displayed as KWIC-concordances together with the corrected forms as well as error annotation.

The following is how we count lexical frequency.

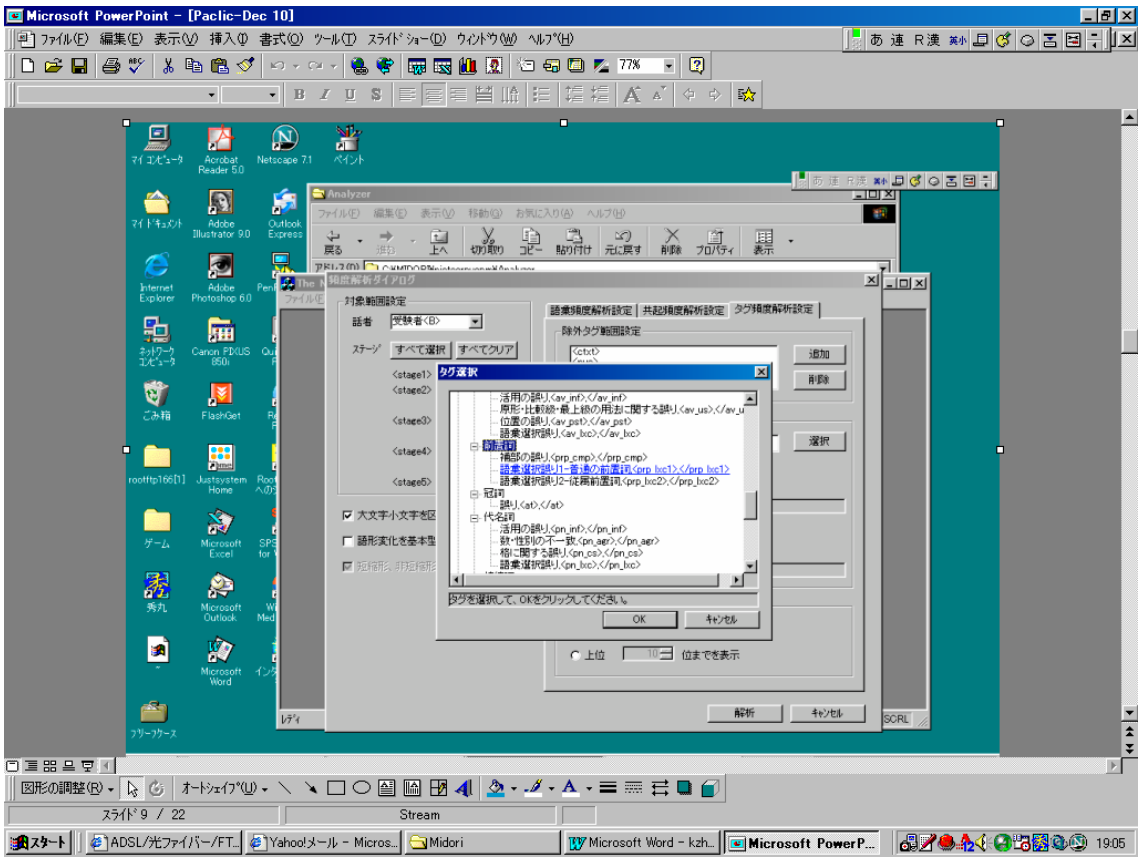
1. Click "analysis tool. exe," after installing the CD. This will bring up the corpus analysis tool.
2. Click the file and then click "concordance and lexical frequency."
3. Choose the learners' proficiency levels that you want to focus on.
4. Choose the speakers, either interviewers or interviewees, or both.
5. Click the "tag frequency."
6. Choose the type of error tag. For example, if you want to look at prepositions, choose an error tag from either <prp_lxc1> or <prp_lxc2>. However, you are not allowed to choose more than one tag.
7. Click the corrected form. This allows you to see incorrect instances as well as respective corrected instances at the same time, when the results are shown. You can also exclude unnecessary tags for the analysis such as fillers.
8. Click "analysis."
9. The results are shown in 4 columns: from the left, error tags, incorrect instance, respective corrected instance and frequency.
10. The result can be sorted by clicking the bar at the top.
11. The results can be stored in alphabetical or frequency order as an excel file.

The instruction shown above is basic and therefore it is suitable for us to glance at it before moving onto the detail. This would be a good start to oversee the error distribution. However, since the error tagging system is complicated for handling the data, we have to be very familiar with how the error tagging system is constructed. For example, the tag <prp_cmp> does not mean prepositions are misused, but it means complements after prepositions are misused. Therefore, errors with the tag <prp_cmp> have to be excluded, in analyzing preposition errors.

Another problem is that error annotation is rather variable. As mentioned earlier, Why-scheme generally describes errors in the following two major ways- error-tag based on a predefined error-hierarchy and narrative - based on natural language descriptions. These two ways have obvious pros and cons in terms of machine and human readability. Sharing and reusing corpora annotated with

different error-typology is difficult. Similarly, with narrative annotation, converting the human understandable descriptions into machine-readable form remains a major challenge.

Essentially, learners' errors are so complex, there is no perfect way to describe them. Nevertheless, although the lack of concrete criteria or established definition is pointed out (see Darwin & Gray 1999), it is more important and constructive to develop ways to facilitate integration and reuse learner corpora.



2.2 Occurrence of three structural error taxonomies

The NICT corpus contains 47 different types of errors as well as 3 structural error taxonomies- omission, addition and misuse. Addition-type means that a grammatical marker is used redundantly where it is not required. Omission-type means that a grammatical marker is missing from where it is obligatorily required. Misuse-type means that a grammatical marker is used inappropriately. Examples are given below.

<Error taxonomy>

Structural type	Error tags	Examples
Addition	<at crr=" ">at</at>	And we were very busy because I was working <prp_lxc1 odr="1" crr="">at</prp_lxc1> full-time at that time (level9)
Omission	<at crr="in"></at>	So <SC>they thin</SC> they <.></.> gotta kind of interest <prp_lxc2 odr="1" crr="in"></prp_lxc2> me. (level6)
Misuse	<at crr="for">to</at>	It's hard <prp_lxc1 odr="1" crr="for">to</prp_lxc1> me to answer the question. (level5)

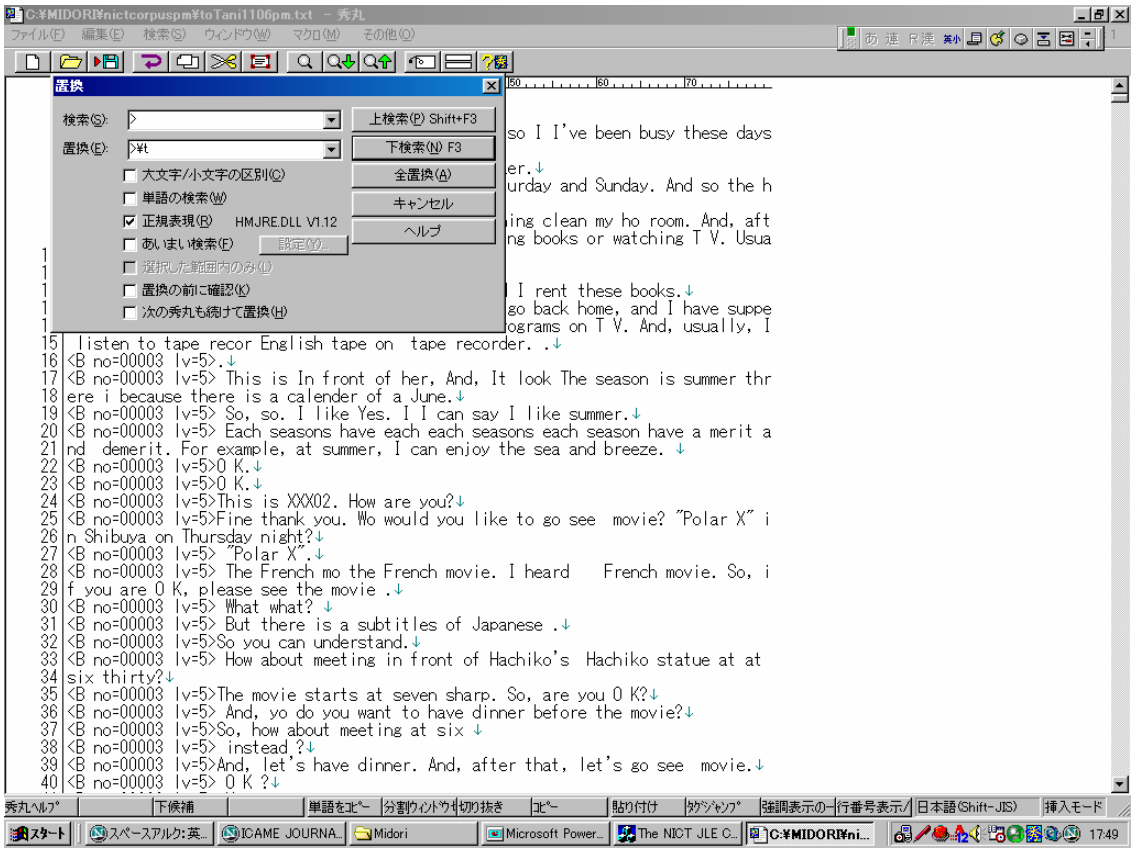
In the same way as described in 2.1, three structural errors are counted.

2.3 Total volume of spoken words

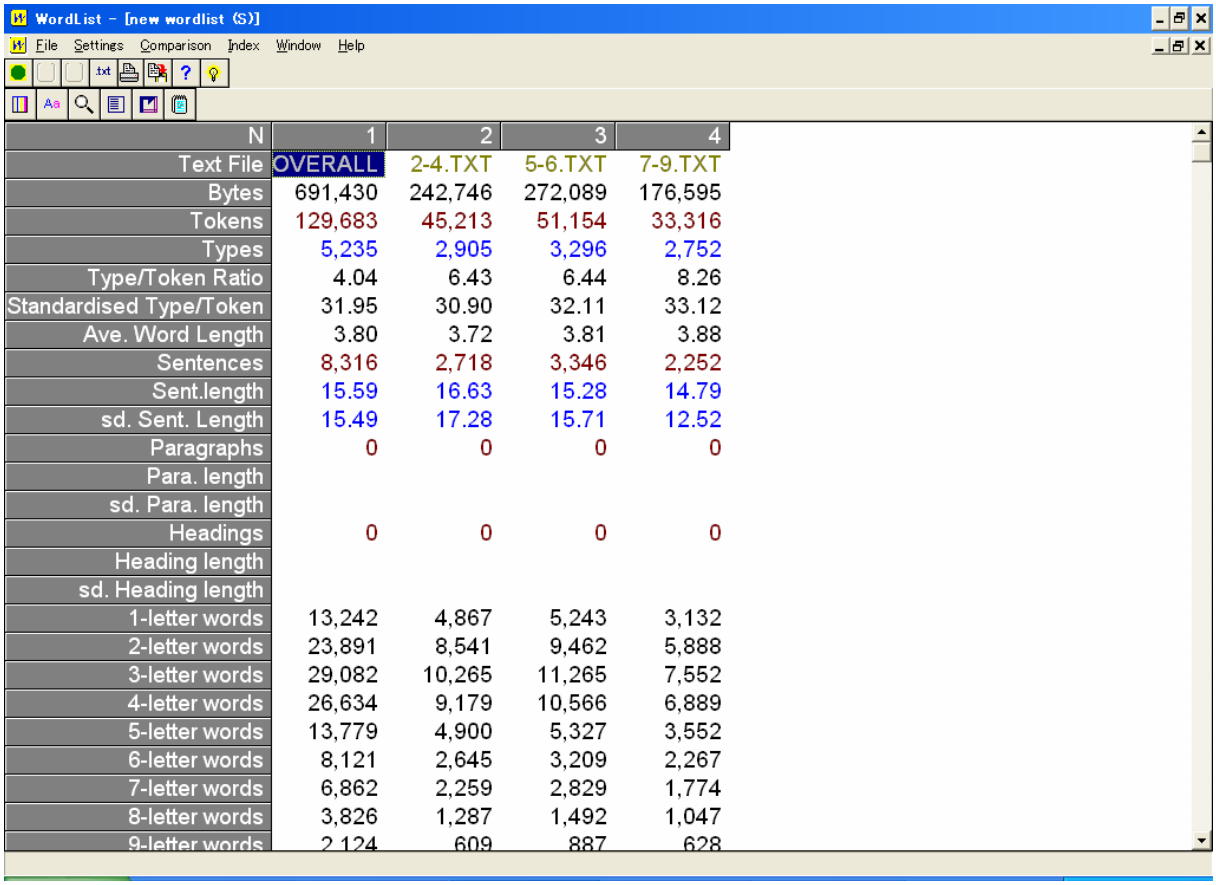
To compare the number of preposition errors among the proficiency levels, first a count is made of the number of occurrences of preposition errors at each level. However, each level contains a different number of error-tagged files - e.g. there is only 1 file in level 1 but 43 files in level 4. In addition, the volume of spoken words normally increases as the proficiency level goes up - e.g. even if learners at level 3 and at level 8 make the same number of errors, learners at higher levels are more likely to speak longer utterances. Therefore, it would be fair to say that just counting the number of errors does not reflect the reality of learners' language usage, and it leads to wrong results.

In order to compare the errors as fairly as possible, the number of errors is converted to "per 1000 words", which is a way of expressing ratios in terms of whole numbers. The NICT JLE corpus analysis tool does not count whole texts. Instead, we use WordSmith *http://www.lexically.net/wordsmith/index.html*, one of the major concordance software, for word counting etc.

However, we can not enter the files into the system of WordSmith, because they are specially converted. In order to allow WordSmith to read the data, we use the XML-like files which are offered by NICT. Using the text files, we categorize 167 files into 9 levels, by checking against the list at the end of the NICT corpus guide book, and we make 9 kinds of text files. Since this learner's corpus contains not only monologs but also dialog, interviewers' parts need to be removed from each file. To this end, we use a text editor, Hidemaru, *http://www.vector.co.jp/soft/win95/writing/se086280.html* and exclude interviewers' parts and also all the tags.



Finally, we obtain only the interviewees' speech without any tags, and we count the number of their spoken words, using WordSmith. Nine kinds of files are divided into 3 level groups: low (level2-4), intermediate (level 5-6) and high (level 7-9). Level 1 is not included for the analysis since there is only one error tagged file.



N	1	2	3	4
Text File	OVERALL	2-4.TXT	5-6.TXT	7-9.TXT
Bytes	691,430	242,746	272,089	176,595
Tokens	129,683	45,213	51,154	33,316
Types	5,235	2,905	3,296	2,752
Type/Token Ratio	4.04	6.43	6.44	8.26
Standardised Type/Token	31.95	30.90	32.11	33.12
Ave. Word Length	3.80	3.72	3.81	3.88
Sentences	8,316	2,718	3,346	2,252
Sent.length	15.59	16.63	15.28	14.79
sd. Sent. Length	15.49	17.28	15.71	12.52
Paragraphs	0	0	0	0
Para. length				
sd. Para. length				
Headings	0	0	0	0
Heading length				
sd. Heading length				
1-letter words	13,242	4,867	5,243	3,132
2-letter words	23,891	8,541	9,462	5,888
3-letter words	29,082	10,265	11,265	7,552
4-letter words	26,634	9,179	10,566	6,889
5-letter words	13,779	4,900	5,327	3,552
6-letter words	8,121	2,645	3,209	2,267
7-letter words	6,862	2,259	2,829	1,774
8-letter words	3,826	1,287	1,492	1,047
9-letter words	2,124	609	887	628

3 Results

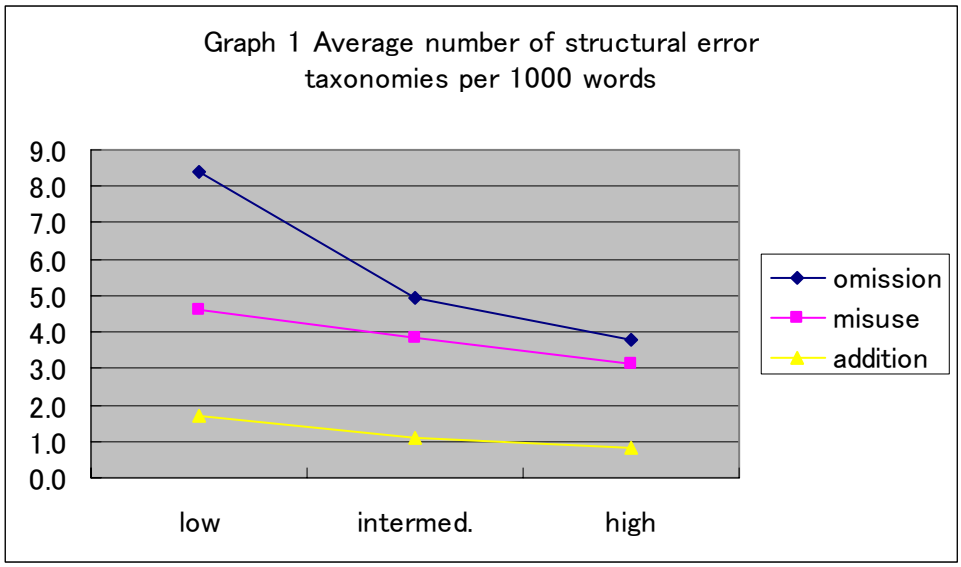
3.1 Qualitative analysis

Table 1 gives the number of participants, the total volume of spoken words, the total of preposition errors and the error ratios per 1000 words in 3 level groups (low, intermediate and high). The right side of the table shows that the error ratio decreases as the proficiency levels go up. In this study, we consider it safe to combine the number of errors of <prp_lxc1> and <prp_lxc2> after counting them respectively.

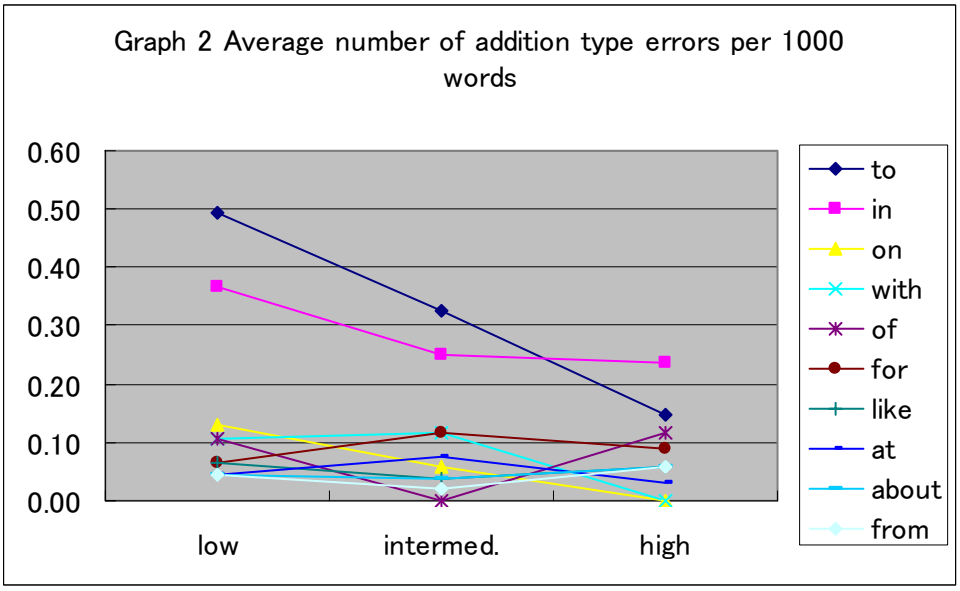
Table 1 The number of occurrences of preposition errors

Level	Number of participants	Total of preposition errors	Amount of spoken words	Errors per 1000 words
Low	78	858	45213	19.0
Intermed.	58	570	51154	11.1
High	30	285	33316	8.6

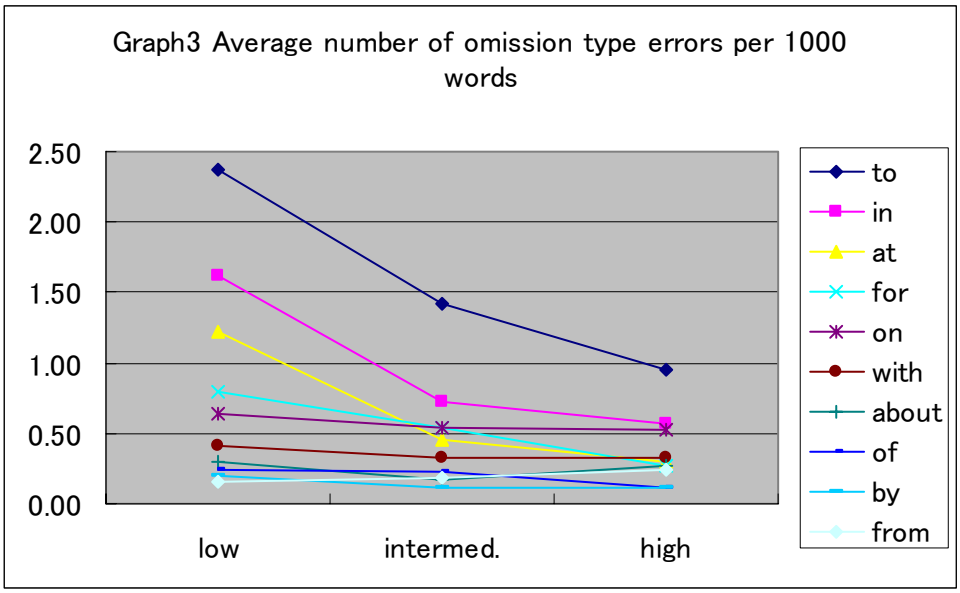
Graph 1 shows the error rates of prepositions in obligatory contexts according to 3 structural error taxonomies: omission, addition and misuse. The number on the right in graph 1 is an error ratio per 1000 words. The ratios of all error types decrease as the proficiency levels go up. In particular, the ratios of omission errors dramatically decrease between low and intermediate. Interestingly, they still remain high, in all levels, compared to the other 2 types.



We compute the number of instances of addition and omission of each preposition. Graph 2 and 3 summarize the results.



Graphs 2 and 3 show that learners typically use words encoding spatio-temporal association and inter-personal relationship such as “to”, “in”, “at” and “with”. Interestingly, both graphs show that error ratio of the top 2 prepositions “to” and “in,” dramatically decreases as the proficiency goes up. The ratio of those two prepositions is very high compared to others. We focus on them in the next section.



3.2 Quantitative analysis

We examine the two prepositions "to" and "in" from the point of view of error structural types, in particular, addition-type and omission-type. Since misuse-type has various errors involved and is difficult to categorize, here we focus on addition-type and omission-type and consider possible L1 transfer effect.

In addition-type, forms of "prepositions + adjectives" in utterances are found in many cases. Examples of this type of utterance are given in. Tags for fillers and repetitions are eliminated.

(1) And I went to I went <prp_lxc1 odr="1" crr="">***to***</prp_lxc1> there with my elder brother.
(2) And my sister-in-law uum um who lives <prp_lxc1 odr="1" crr="">***in***</prp_lxc1> next-door
(3) I met er my father <prp_lxc1 odr="1" crr="">***in***</prp_lxc1> this morning.

At first glance, it seems that the L2 learners do not understand the usage of prepositions, but the usage like "go to there" and "live in next-door" show that they do understand "go to + GOAL" and "live in + PLACE" as a higher level knowledge, and those words normally collocate together in chunks. In fact, these overgeneration utterances account for almost all the cases.

On the other hand, the L2 learners also omit prepositions. In the following examples (4) – (6), omission is found after verbs such as "go", "come," "live" and "be", and most semantic meanings are related to spatio-temporal relationships.

(4) I want to be <prp_lxc1 odr="1" crr="***in***"></prp_lxc1> New York urr early,
(5) <at odr="1" crr="the">this</at> test starts ee <prp_lxc1 odr="2" crr="***in***"></prp_lxc1> July
(6) if possible, I want to go <prp_lxc1 odr="1" crr="***to***"></prp_lxc1> Whistler.

These examples of both addition-type and omission-type interestingly show the factors of possible L1 transfer effect. Take (1) as an example. "there" in English has meanings of both "preposition" and "place", while Japanese "so ko" is case free, and it has meaning when followed by case marker – e.g. "soko e", "soko ni", "soko wo" etc. Therefore we assume that learners put "to" before "there", because they believe that "soko" is the counterpart of "there". On the other hand, in (6), "go" in English only has meaning when followed by prepositions as go to/ go for/ go along with/go in for/ go off/ go on/ go out/ go over etc, while in Japanese, "iku" (=go) has meaning without case markers. That is in Japanese, verbs inherently have argument. More examples are given in (7) and (8).

(7) eki (e / ni) iku
station to go
<GOAL>

(8) michi (wo) iku
road accusative marker go
<PASS>

We can see that Japanese allows us to omit the case markers, and the nouns ("eki" in the case of (7) and "michi" in the case of (8)) automatically determine the meaning of "iku." In Talmy (1985), lexicalization patterns in typological difference are discussed, and we assume that this difference would be a result of L1 effect, but we leave this issue for future work.

3.3 Further direction to reuse leaner corpora

Before concluding our study, further direction is given to facilitate integration and reuse learner corpora.

Variable tagging
A) <prp+lxc1> and <prp+lxc2>
(9) so we went <prp_lxc2 odr="1" crr="***to***"></prp_lxc2> the restaurant together. (level 6)
(10) if possible, I want to go <prp_lxc1 odr="1" crr="***to***"></prp_lxc1> Whistler. (level 4)
B) omission type and addition type
(11) Then, <prp_lxc1 odr="1" crr="***in***"></prp_lxc1> this season, urr he is the captain of <at odr="2" crr="the"></at> te team.
(12) I met er my father <prp_lxc1 odr="1" crr="">***in***</prp_lxc1> this morning.

Error correction by verb changes
(13) we <v_lxc odr="1" crr="talked">discussed</v_lxc> <prp_lxc2 odr="2" crr="***about***"></prp_lxc2> the menu.

In (9) and (10), the errors about "go to PLACE" are found in both <prp_lxc1> and <prp_lxc2>. They are supposed to be tagged in the same way, but in (9) "to" is understood as a part of phrasal verb "go to," but in (10) "to" is understood as a part of prepositional phrase "to + PLACE". Since no concrete definition has been established in research and pedagogy, it is not surprising that a certain degree of arbitrariness is involved. In (11), "this season" which is uttered by the interviewee is actually correct, and the annotator may misunderstand it as an error, while in (12), error annotation is correct. Finally, (13) seems to include a preposition error at a glance, but closer inspection shows that a preposition error is a result of changing the verb "discuss" to "talk". Although there are some points which need to be improved it is more important to identify preposition errors.

4 Conclusion

The present study aims firstly to show what the NICT JLE corpus analysis tool can do, and show how various other tools such as Hidemaru and WordSmith can be integrated to analyse in detail. Secondly, we show that close analysis of error tagged corpus would give us objective evidence for particular errors, and this would lead to further study – such as what would be a factor of learners' errors. Easy and convenient analysis tools seem to be a miracle breakthrough in computer technology, but they have limitation in dealing with data. Since learners' errors are so various and complex, a false conclusion can easily be derived from just counting the data without looking at details. It is important to facilitate integration and reuse learner corpora.

As a future direction in second language acquisition research, we need to develop an objective method to characterize L1 transfer. We have an assumption that back-translation of each utterance gives us

evidence to characterize or extract L1 translation objectively. Bi-lingual aligned corpus and machine translation software will be a foundation to develop such a method. We are now continuing to examine this assumption. We also need to standardize a shared collocation dictionary. A list of phrasal verbs should be designed for use in the form of "lexicon" (i.e. JACET 8000, a word list with 5 different proficiency levels, especially for Japanese English learners for pedagogic purposes). In order to do this, it is necessary to share the corpus, which determines if a certain collocation is a phrasal verb, or not.

References

Darwin, C. M. and L. S. Gray. 1999. Going after the phrasal verb: An alternative approach to classification. *TESOL Quarterly* vol.33. 65-83.

Talmy, L. 1985. Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen, ed., *Grammatical Categories and the Lexicon. Language, Typology and Syntactic Description 3* (pp.57–149). Cambridge University Press.

Tono, Y., Kaneko, T., Isahara, H., Saiga, T. and E. Izumi. 2001. The Standard Speaking Test Corpus: A 1 million-word spoken corpus of Japanese learners of English and its implications for L2 lexicography. In S. Lee, ed., *ASIALEX 2001 Proceedings: Asian Bilingualism and the Dictionary*.(pp. 257-262). The Second Asialex International Congress, August 8-10, 2001, Yonsei University, Korea.

Tono, Y. 2004. The NICT JLE Corpus ni miru eigogakushusha no happyogoi no shiyojyokyo. In E. Izumi, K. Uchimoto and H. Isahara Nihonjin, eds., *1200nin no eigo speaking corpus* (pp.96-112). ALC.